COMENIUS UNIVERSITY IN BRATISLAVA
FACULTY OF MATHEMATICS PHYSICS AND INFORMATICS

CONTEXTUALIZED LANGUAGE MODEL-BASED NAMED ENTITY RECOGNITION IN SLOVAK TEXTS

Diploma thesis

2021 Bc. Dávid Šuba

COMENIUS UNIVERSITY IN BRATISLAVA
FACULTY OF MATHEMATICS PHYSICS AND INFORMATICS

CONTEXTUALIZED LANGUAGE MODEL-BASED

Diploma thesis

Study program:	Applied informatics
Branch of study:	2511 Applied informatics
Supervisor:	Mgr. Endre Hamerlik
Consultant:	Mgr. Marek Šuppa

Bratislava, 2021 Bc. Dávid Šuba

I hereby declare that I have written this thesis by myself, only with help of referenced literature, under the careful supervision of my thesis advisor.

....

Bratislava, 2021 Bc. Dávid Šuba

Acknowledgement

I want to thank....

Abstract

Named Entity Recognition (NER) is one of the fundamental tasks in Natural Language Processing (NLP), with English state-of-the-art approaches generally utilizing neural models. The currently available NER classifiers for Slovak texts are either rule- and vocabulary-based systems or employ multilingual Contextualized Language Models. Both of these show poor performance compared to the language-specific Deep Contextualized Language Models, even in low-resource languages, such as Slovak.

Keywords: named entity recognition, natural language processing, deep learning

Abstrakt

Rozpoznávanie pomenovaných entít (NER) je jedna zo základných úloh v spracovaní prirodzeného jazyka, kde najlepšie modely sú založené na neurónových sieťach. Aktuálne dostupné NER klasifikátory sú založené buď na pravidlách a slovnej zásobe alebo multi-jazykových kontextualizovaných jazykových modeloch. Obidva prístupy nedosahujú úroveň pre jazykovo špecifické hlboké kontextualizované jazykové modely dokonca ani v jazykoch s málo zdrojmi ako slovenčina.

Kľúčové slová: rozpoznávanie pomenovaných entít, spracovanie prirodzeného jazyka, hlboké učenie

Contents

1 Intro **1**

Chapter 1

Intro

Named Entity Recognition (NER) is one of the fundamental tasks in Natural Language Processing (NLP), with English state-of-the-art approaches generally utilizing neural models. The currently available NER classifiers for Slovak texts are either rule- and vocabulary-based systems or employ multilingual Contextualized Language Models. Both of these show poor performance compared to the language-specific Deep Contextualized Language Models, even in low-resource languages, such as Slovak.

Bibliography

[RLC19] Afshin Rahimi, Yuan Li, and Trevor Cohn. Massively multilingual transfer for ner, 2019.

[SJ21] Marek Suppa and Ondrej Jariabka. Benchmarking pre-trained language models for multilingual NER: TraSpaS at the BSNLP2021 shared task. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 105–114, Kiyv, Ukraine, April 2021. Association for Computational Linguistics.

List of Figures